Projekt A - Diamantpris-prediktion

Piere Ventura Cruz, Axel Johansson & Simon Jorstedt

2021-09-27

Sammanfattning

I den här rapporten ska vi prediktera diamanters försäljningspris med hjälp av förklarningsvariablerna vikt (carat), färg, klarhet och organisationen som certifierat diamanten. När vi plottar pris mot vikt märker vi att vår regressionslinje passar något bra till datan men att vikten verkar ha ett kvadratiskt beteende. Vi märker också att vi måste logaritmera priset. Vi studerar multikollinearitet med hjälp av VIF och använder oss utav Forward selection samt Backwards elimination för att få fram våra modeller.

Introduktion

Det primära syftet med denna rapport är att studera, och att försöka prediktera diamanters säljpris. På vägen kommer vi ta en titt på skillnader mellan de certifikerande institutioner som bedömmer diamanterna. Vi kommer sedan gå vidare till det primära syftet genom att studera de tillgängliga förklaringsvariablerna, hur de relaterar till varandra och hur de påverkar priset..

Data

Datamaterialet består av egenskaperna Carat, färg (*Color purity*), klarhet (*Clarity*), Organisation, och pris (Singaporianska dollar S$) för 308 diamanter. I brist på information om var priset kommer från, antar vi att priset och diamanterna i datamängden är tagna från 308 individuella oberoende försäljningar. Vanligtvis brukar även skurningen *Cut* av en diamant vara intressant, men den variabeln ingår inte datamaterialet.

Carat (ej Karat) är ett viktmått på diamanter, motsvarande 0.2g. färg är ett mått på en diamants renhet, som anges i de alfabetiska kategorierna D, E, F, G, H... i nedstigande led. För att kunna genomföra statistisk analys har kategorierna översatts till sifferbetygen 6-1. Kategorierna är visserligen D-Z, men endast kategorierna D-H återfinns i datamaterialet. Klarhet anger förekomsten av skrapningar i diamanten. Kategorierna IF, VVS1, VVS2, VS1 och VS2 har översatts till siffervärdena 5-1. Organisation anger den oberoende organisation som utvärderat och utfärdat ett certifikat för en specifik diamant. De tre organisationerna som förekommer i datamängden är Gemological Institute of America (GIA), International Gemological Institute (IGI) samt Hoge Raad Voor Diamant (HRD). De kodas med två dummy-variabler, där GIA betraktas som utgångsläget.

Institutionerna

Vi plottar upp de tre instutionerna mot varandra för att se hur de skiljer sig när det kommer till vilka sorters diamanter de värderar. Detta visas i figur 1.

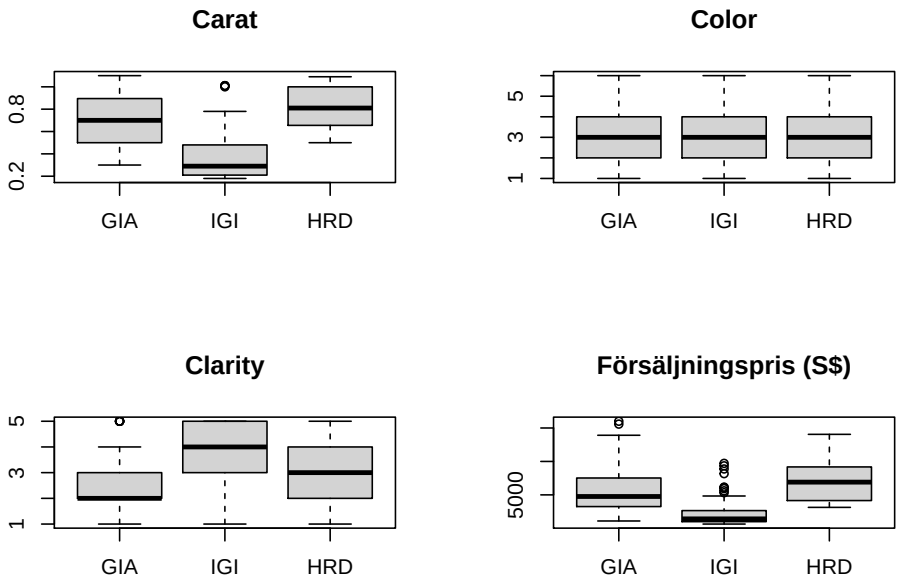

Figure 1: Boxplot för vikt, färg, klarhet och försäljningspris för diamanterna från de tre respektive organisationerna GIA, IGI and HRD.

Tydligen skiljer sig diamanterna från de tre certifikatutfärdarna mycket i Carat, Clarity och slutpris. Anledningen till dessa tydliga skillnader är troligtvis antingen att de olika institutionerna tilldelas olika typer av diamanter att certifiera, eller det något *otroligare* alternativet att institutionerna och därför skulle bedömma en hypotetisk diamant med olika "betyg".

Förklaringsvariabler

Vi skall inleda med att ta en titt på de aktuella förklaringsvariablerna, hur de relaterar till priset, och hur de relaterar till varandra.

Vikt

Vi inleder med att plotta priset mot vikten i figur 2. Vikten är den enda icke-kategoriska variabeln och därför en naturlig utgångspunkt. Vi konstruerar även en enkel reglinje för pris mot vikt i figur 2.

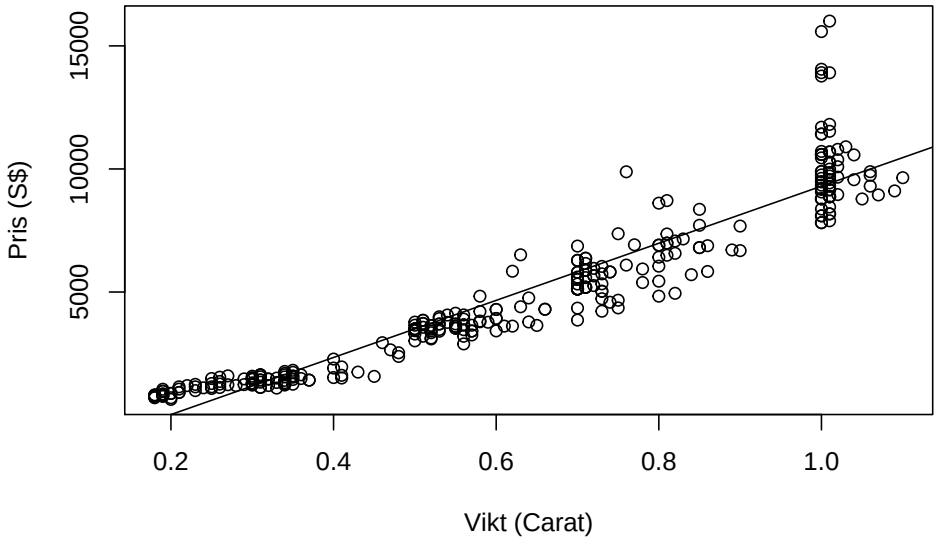

Figure 2: Pris plottad mot vikt, samt enkel regressionslinje.

I figur 2 ser vi en något sånär bra anpassning, men det är tydligt att något är fel. Priset verkar uppvisa ett icke-linjärt beteende, möjligen kvadratiskt. För att undersöka detta ytterligare plottar vi anpassningens residualer i figur 3.

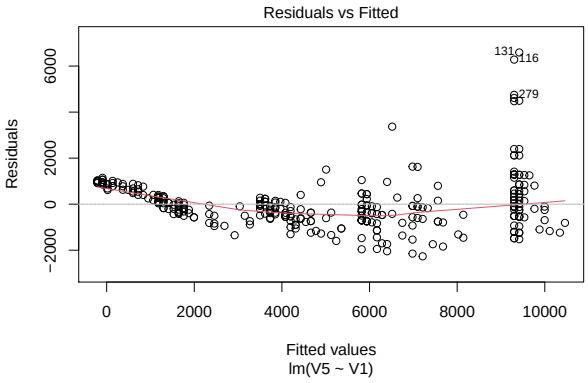

Figure 3: Residualer för den enkla linjära anpassningen i figur 2: lm(pris = vikt).

I figur 3 är det möjligen kvadratiska beteendet ännu tydligare. För att bättre beskriva sambandet mellan vikt och pris lägger vi därför till en term i den enkla modellen motsvarande vikten i kvadrat. I figur 4 ser vi residualerna för den modellen.

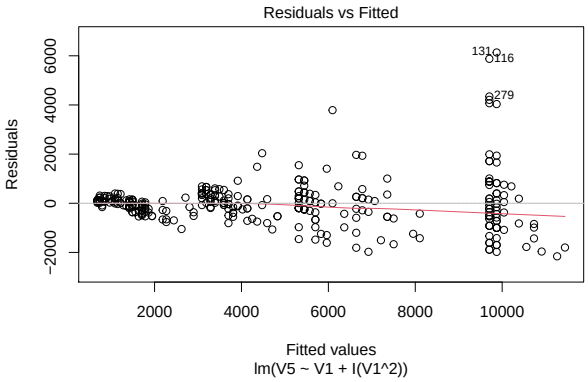

Figure 4: Residualer för en anpassad modell med vikt, samt vikt i kvadrat som förklaringsvariabler för priset. Modellen: lm(pris = vikt + vikt^2).

I figur 4 är residualerna utjämnade längs 0-linjen, men det råder fortfarande heteroskedecitet. För större vikter ökar alltså vårt anpassningsfel, medan felet är lågt för små vikter. Vi försöker åtgärda detta genom att logaritmera responsvariabeln pris. Vi anpassar alltså en model med vikt, samt kvadrerad vikt som förklaringsvariabler, och det logaritmerade priset som responsvariabel. Vi använder den naturliga logaritmen. Resultatet av detta framställs i figur 5.

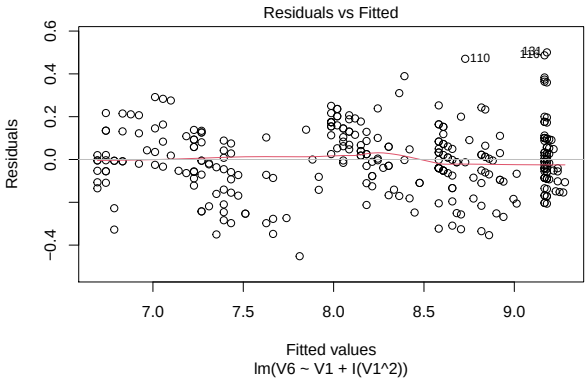

Figure 5: Residualer för den anpassade modellen med förklaringsvariablerna vikt och kvadrerad vikt, och logaritmerat pris som responsvariabel. Modellen: lm(log[pris] = vikt + vikt^2).

Residualplotten i figur 5 är mycket jämnare än i figur 4. Tydligen har vi lyckats åtgärda heteroskedeciteten. Vi skall även i fortsättningen betrakta det logaritmerade priset som vår responsvariabel.

Färg och Klarhet

Vi skall nu gå vidare med att studera färg och klarhet. Vi fortsätter med att betrakta det logaritmerade priset som responsvariabel för att kunna föra samman förklaringsvariablerna i samma modell senare. I figur Ä ser vi det logaritmerade priset plottat mot färg respektive klarhet, samt motsvarande enkla regressionslinjer.

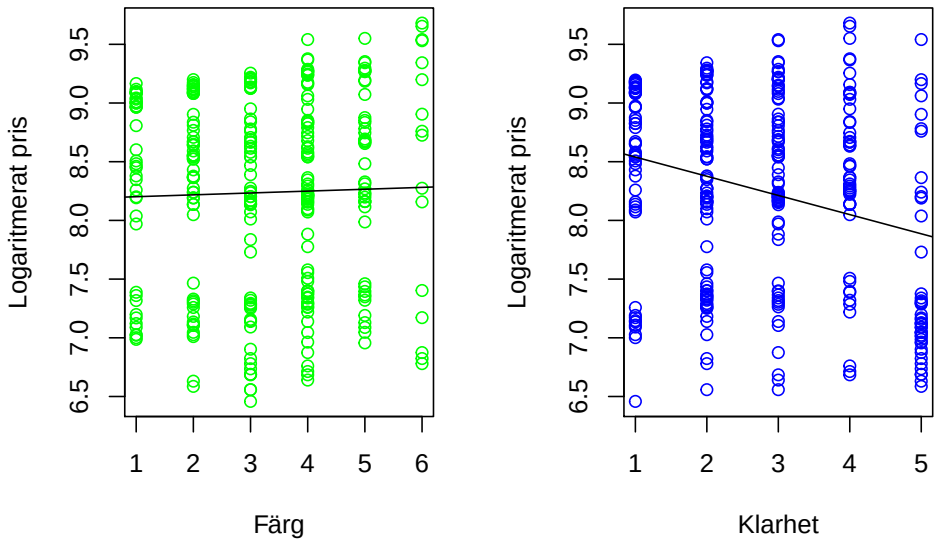

Figure 6: Logaritmerat pris plottat mot färg respektive klarhet, samt enkla reglinjer.

Det vi ser i figur 6 är att det tydligen råder ett negativt samband mellan klarhet och logaritmerat pris. Men det är missvisande. Som tidigare konstaterats är institutionen IGI överrepresenterad bland diamanter med låg vikt, såväl som hög clarity. Eftersom vikt verkar ha en mer betydande effekt på priset än klarhet så "maskeras" därför effekten av klarheten och framställs som negativ. I själva verket är klarheten *troligtvis* associerat med högre pris (och därför även logaritmerat pris) i en diamant-population som är oberoende från institutionerna.

I figur 6 ser vi även att det verkar råda en viss svag (positiv) korrelation mellan färg och logaritmerat pris. Det resultatet har vi mer tilltro till än clarity-beteendet, eftersom vi i figur 1 sett att color verkar vara jämt fördelat mellan institutionerna.

Institutionerna

De tre institutionerna har kodats med två dummy-variabler för att kunna inkluderas i analys och modellkonstruktion. Vi väljer GIA till grund-läge, eftersom diamanterna därifrån verkar någorlunda jämt fördelade över egenskaperna vikt, pris, färg och klarhet.

Multikollinearitet

Vi skall nu studera eventuell multikollinearitet i förklaringsvariablerna vikt, kvadrerad vikt, färg och klarhet. Vi gör detta genom att beräkna de respektive VIF-värdena. Det är oklart hur ett VIF-värde för Dummyvariablerna skulle beräknas (på ett meningsfullt sätt), så det utelämnar vi. Men däremot tas dummyvariablerna med i beräkningen av de övriga VIF-värdena.

Gör vi detta så får vi resultaten $VIF_{vikt} \approx 38.56$, $VIF_{färg} \approx 1.02$, $VIF_{klarhet} \approx 1.39$, samt $VIF_{vikt^2} \approx 34.47$. Det framgår att vikt och vikt i kvadrat är multikollineära, vilket är fullt rimligt. Färg och klarhet är inte signifikanta. Vi konstruerar tre modeller med olika kombinationer av vikt och kvadrerad vikt som förklaringsvariabler, och det logaritmerade priset som responsvariabel och jämför dem. Följande är summary av dessa tre anpassade modeller.

```
##
## Call:
## lm(formula = V6 ~ V1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55486 -0.16271 -0.00869  0.15522  0.59431
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.44488    0.02938  219.40   <2e-16 ***
## V1           2.84155    0.04264   66.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2071 on 306 degrees of freedom
## Multiple R-squared:  0.9355, Adjusted R-squared:  0.9353
## F-statistic:  4441 on 1 and 306 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = V6 ~ I(V1^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86905 -0.24726 -0.01934  0.28005  0.74547
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.24362    0.03056  237.00   <2e-16 ***
## I(V1^2)      2.09434    0.05148   40.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3222 on 306 degrees of freedom
## Multiple R-squared:  0.8439, Adjusted R-squared:  0.8434
## F-statistic:  1655 on 1 and 306 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = V6 ~ V1 + I(V1^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45187 -0.08858 -0.00441  0.09685  0.50045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.7806     0.0483  119.68   <2e-16 ***
## V1            5.4368     0.1709   31.81   <2e-16 ***
## I(V1^2)      -2.0501     0.1326  -15.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1553 on 305 degrees of freedom
## Multiple R-squared:  0.9639, Adjusted R-squared:  0.9636
## F-statistic:  4066 on 2 and 305 DF,  p-value: < 2.2e-16
```

Vi kan återigen se att vikten ensam ger en mycket hög (justerad) förklaringsgrad $R_v^2 \approx 0.94$, medan kvadrerad vikt ger en något lägre: $R_{v^2}^2 \approx 0.84$. Tillsammans blir modellen bara snäppet bättre ($R_{v+v^2}^2 \approx 0.96$) än när vi endast inkluderade vikten. Detta skulle kunna leda oss till att avfärda den kvadrerade vikten, men vi har tidigare konstaterat förekomsten av det kvadratiska sambandet och väljer därför att ändå behålla den kvadrerade vikten.

Modellval och prediktion

Vi skall nu studera och försöka konstruera en bra modell som beskriver datan tillfredställande, och som förhoppningsvis även har bra prediktiva förmågor.

Forward selection

Vi börjar med att genomföra forward selection-metoden, som vi anpassat för situationen enligt följande: Vi börjar med en konstant modell, och lägger till den förklaringsvariabel som ger det högsta (justerade) R^2-värdet, om den förklaringsvariabels parameter i den modellen är signifikant på 95%-nivån. Vi fortsätter på detta sätt, tills vi uppnått ett (justerat) R^2-värde på minst 0.99.

Anledningarna till denna utformning och att vi valt att jämföra R^2-värdena är bland annat att två p-värden erhålls som modelljämförelsemått när V7 och V8 läggs till, vilket är svårtolkat. Genom att ändå lägga till kravet att tillagda förklaringsvariabler ska ha signifikant parameter tar vi hänsyn till eventuellt osäkra parametrar.

När denna metod genomförs inkluderas vikt, sedan kvadrerad vikt, färg, och slutligen klarhet i en modell med ett (justerat) R^2-värde på ca 0.995. Denna modell kallar vi *VV2FK*.

Backwards elimination

Vår backwards elimination-metod specifierar vi enligt följande: Först inkluderas alla förklaringsvariabler. Det vill säga vikt, kvadrerad vikt, färg, klarhet samt Dummyvariablerna. Vi utesluter den förklaringsvariabel som motsvarar minst minskning i (justerade) R^2-värdet när den tas bort. Vi fortsätter utesluta förklaringsvariabler på det här sättet, så länge den kvarvarande modellen minst har (justerat) R^2-värde 0.95. Vi kräver även att samtliga parametrar i en kvarvarande modell skall vara signifikanta på 95%-nivån.

Med denna specifikation har den inledande modellen ett justerat R^2-värde på ca 0.995. Vi utesluter först Dummyvariablerna, sedan färg, och slutligen klarhet. Vi kan inte dessutom utesluta antingen vikt, eller kvadrerad vikt utan att modellens (justerade) R^2-värde understiger 0.95. Denna modell kallar vi *VV2*.

Utvärdering

Innan vi utvärderar modellerna *VV2FK* och *VV2*, så behöver vi återvända till responsvariabeln. Alla våra anpassningar har varit av det logaritmerade priset, varför vi nu behöver lösa ut priset P genom att höja båda leden med e. Vi betecknar vikt med V, kvadrerad vikt med V^2, färg med F, samt klarhet med K. Modell *VV2FK* blir då

$$P_i = exp(\alpha + \beta_1 V_i + \beta_2 V_i^2 + \beta_3 F_i + \beta_4 K_i) = e^{\alpha} \cdot e^{\beta_1 V_i} \cdot e^{\beta_2 V_i^2} \cdot e^{\beta_3 F_i} \cdot e^{\beta_4 K_i}$$

På liknande sätt får vi att modell *VV2* blir

$$P_i = exp(\alpha + \beta_1 V_i + \beta_2 V_i^2) = e^{\alpha} \cdot e^{\beta_1 V_i} \cdot e^{\beta_2 V_i^2}$$

Notera att parametrarna i de två modellerna är olika, men notationen återanvänds för att undvika onödigt komplex notation. Exempelvis gäller att β_1 från *VV2FK* inte är lika med β_1 från modell VV2.

MSEP

Nu när modellerna är konstruerade vill vi utvärdera deras prediktionsförmågor. För enkelhetens skull beräknar vi MSEP-värdena för modellversionerna med det logaritmerade priset som respons. Det förändrar inte resultatet, eftersom vi ju bara är intresserade av vilken modell som har det största MSEP-värdet.

Vi får då MSEP-värdena 0.0036155 (*VV2FK*) och 0.0243369 (*VV2*). Det lägre värdet för *VV2FK* säger oss att *VV2FK* är bättre på att prediktera ny data (av ungefärlig samma typ som i datamängden) än *VV2*. Men denna skillnad är mycket liten, och beroende på syftet är det mycket möjligt att den är försumbar.

Resultat

Den bästa modellen som vi kommit fram till är slutligen *VV2FK* som kan avrundas till följande:

$$P_i = e^{5.14} \cdot e^{5.72 V_i} \cdot e^{-2.14 V_i^2} \cdot e^{0.09 F_i} \cdot e^{0.08 K_i}$$

Det skall återigen poängteras att även modell *VV2* var en god anpassning till data, och den kan avrundas till följande:

$$P_i = e^{5.78} \cdot e^{5.44 V_i} \cdot e^{-2.05 V_i^2}$$

VV2 har dessutom fördelen att kunna plottas meningsfullt i två dimensioner. Detta ses i figur 7.

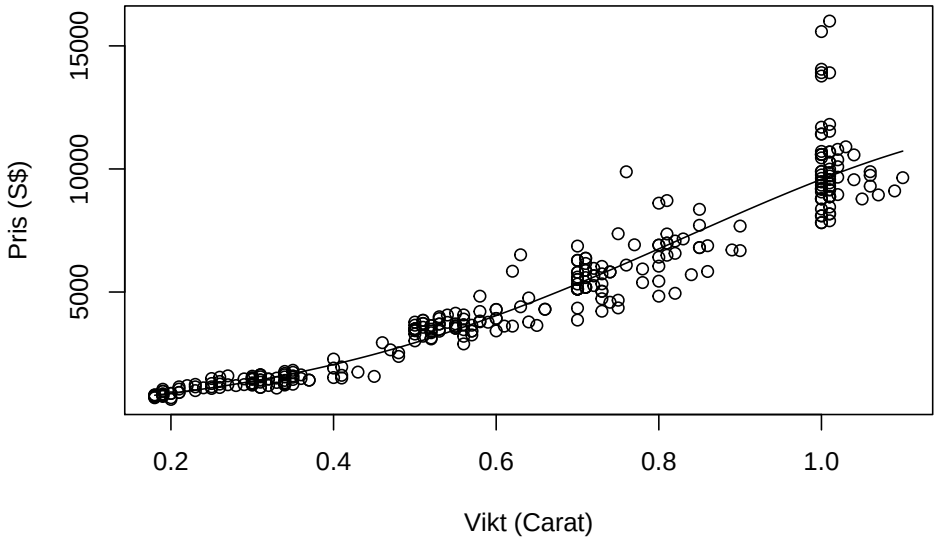

Figure 7: Pris plottad mot vikt, tillsammans med kurva för modell VV2.